

Can We Talk Models Into Seeing The World Differently?

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos,

M. Jehanzeb Mirza ,Margret Keuper, Janis Keuper. ICLR 2025

3/12/25 Group 10, Presented by:

Mehrab Mehdi Islam, Michael Miller, Michael Ishak, Lam Nguyen,

Robinson Vasquez Ferrer

CONTENTS

1 Visual Cues in humans & CV

2 How to measure biases

3 VLM cue bias analysis

4 VLM cue bias steering

5 Generalize to other cues

6 Conclusion

7 Future Work

8 References

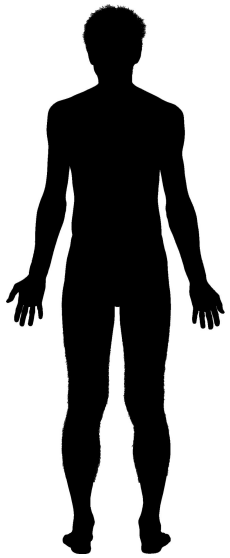
Lam’s slides are first

Can you talk to a model to get it to change its’ perception?

- Text input to a Vision Language Model will change the nature of how a model identifies an image and adds a caption.
- Computer Vision problem combined with an LLM

Visual Cues: *The Texture-Shape Bias in Humans*

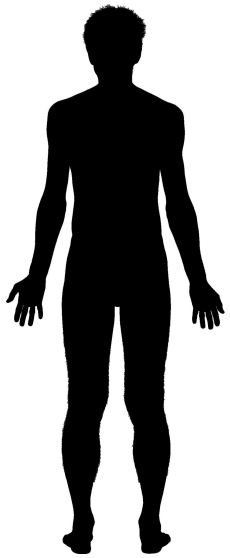
Humans identify objects through the visual cue of *shape.*

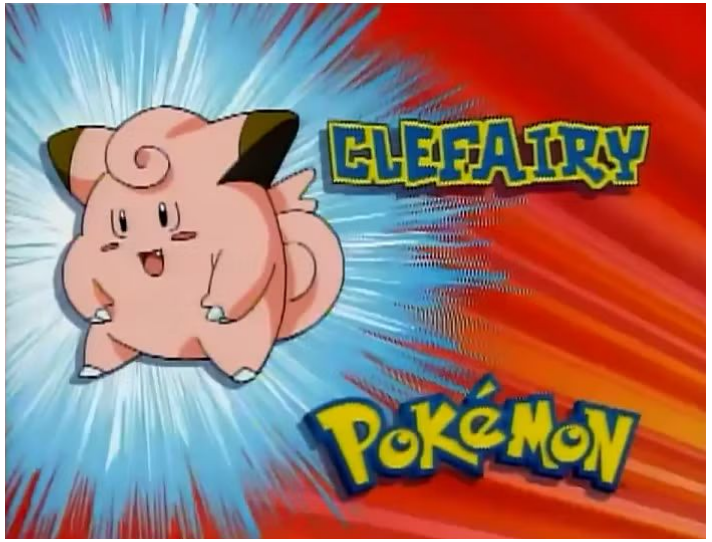




Visual Cues: *The Texture-Shape Bias in Humans*

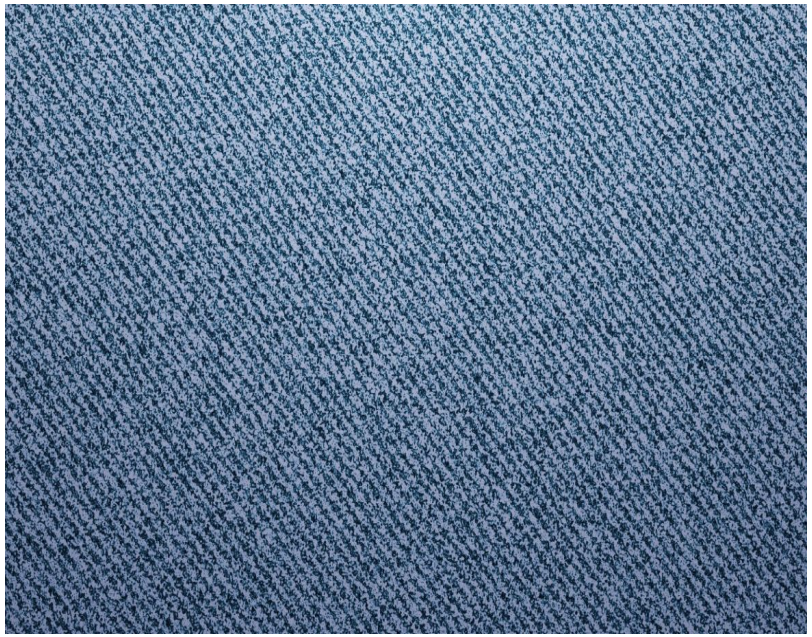
Humans identify objects through the visual cue of *shape.*





Visual Cues: *Texture-Shape Bias in Machine Learning Models*

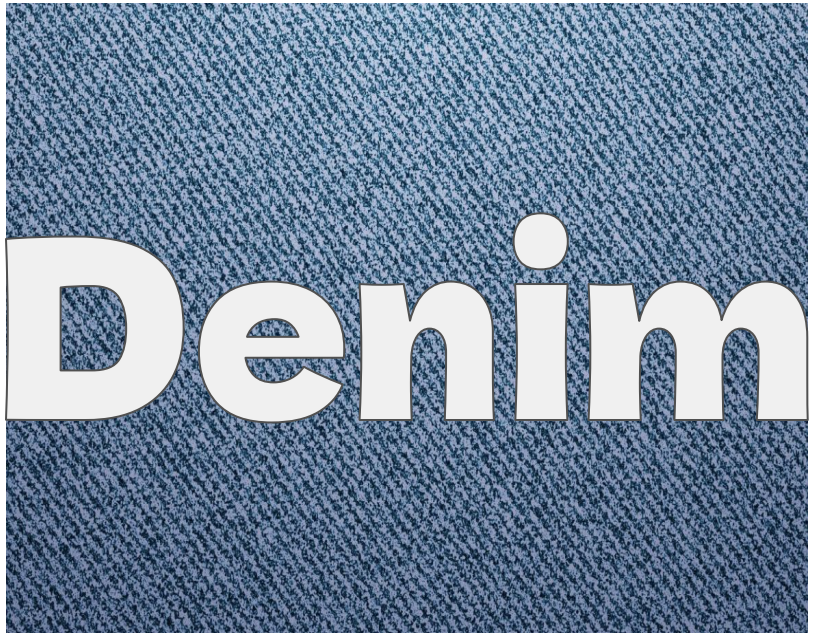
ML models tend to identify an object through the visual cue of *texture.*





Visual Cues: *Texture-Shape Bias in Machine Learning Models*

ML models tend to identify an object through the visual cue of *texture.*





The Texture-Shape Bias in CV Models

Cue bias: shape and the texture are mismatched



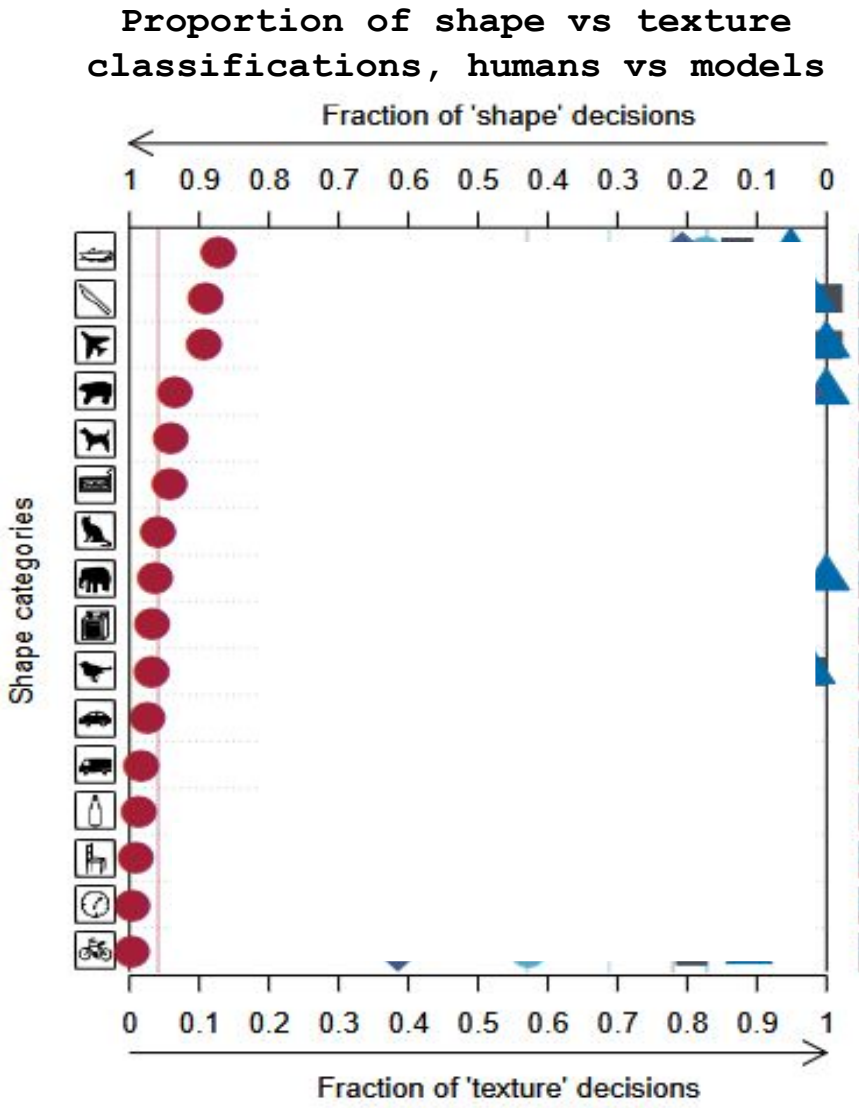
💁‍♀️: Car!

🤖: Clock!

Source: Figure 4 of Geirhos et al, 2019

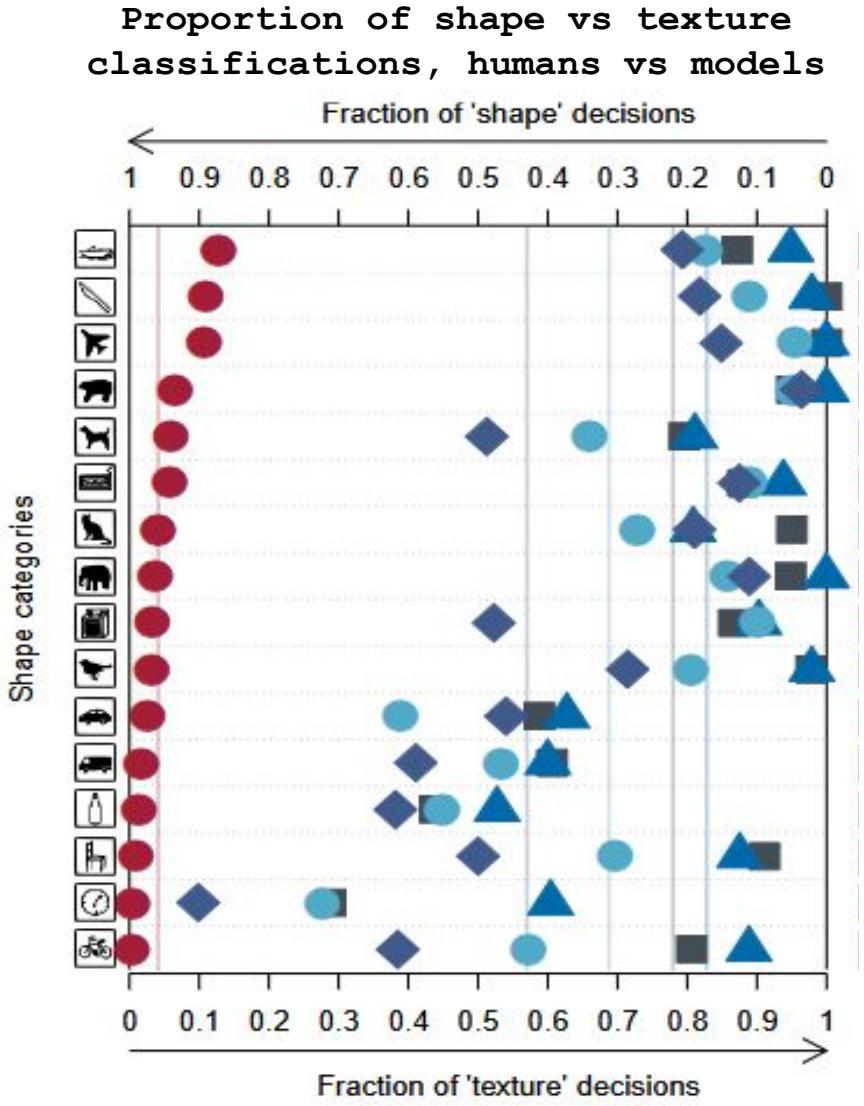
The Texture-Shape Bias in CV Models

97 Humans



Source: Figure 4 of Geirhos et al, 2019

The Texture-Shape Bias in CV Models



97 Humans
AlexNet
GoogLeNet
VGG-16
ResNet-50

Source: Figure 4 of Geirhos et al, 2019

Can We Steer Models' Bias Towards Humans' Bias?

What Might this Imply?

- Through prompts, one would be able to actually change the structure of a machine learning model without having to rewrite the code by hand or retrain the model
- Possibly an Visual Language Model that can rewrite itself on a more fundamental level then just changing the weights of the parameters?

Where do the biases come from?

Related Work - How Do We Measure Texture/Shape Bias?

Cue Conflict Dataset - 1,280 images where shape and texture are swapped







(Picture source: Geirhos et al.. 2019)

Related Work - How Do We Measure Texture/Shape Bias?

$$\textit{Cue Accuracy} = \textit{Shape Accuracy} + \textit{Texture Accuracy}$$

$$\textit{Shape Bias} = \textit{Shape Accuracy}/\textit{Cue Accuracy}$$

Methodology - Measuring VLM Cue Biases Through Classification

Visual Question Answering (VQA)

```
"{VQA_INSTRUCTION}
A. airplane
B. bear
C. bicycle
…
P. truck
Answer with the option's
letter from the given
choices directly."
```

Prompt
"Which option best describes the image? [...]" (default)
Default with CLIP-style options
"Describe the object in the image: [...]"
"Describe the object in the image: [...]" with CLIP-style options
Empty instruction (just options)

Source: Table 5, Appendix C.1.

Methodology - Measuring VLM Cue Biases Through Classification

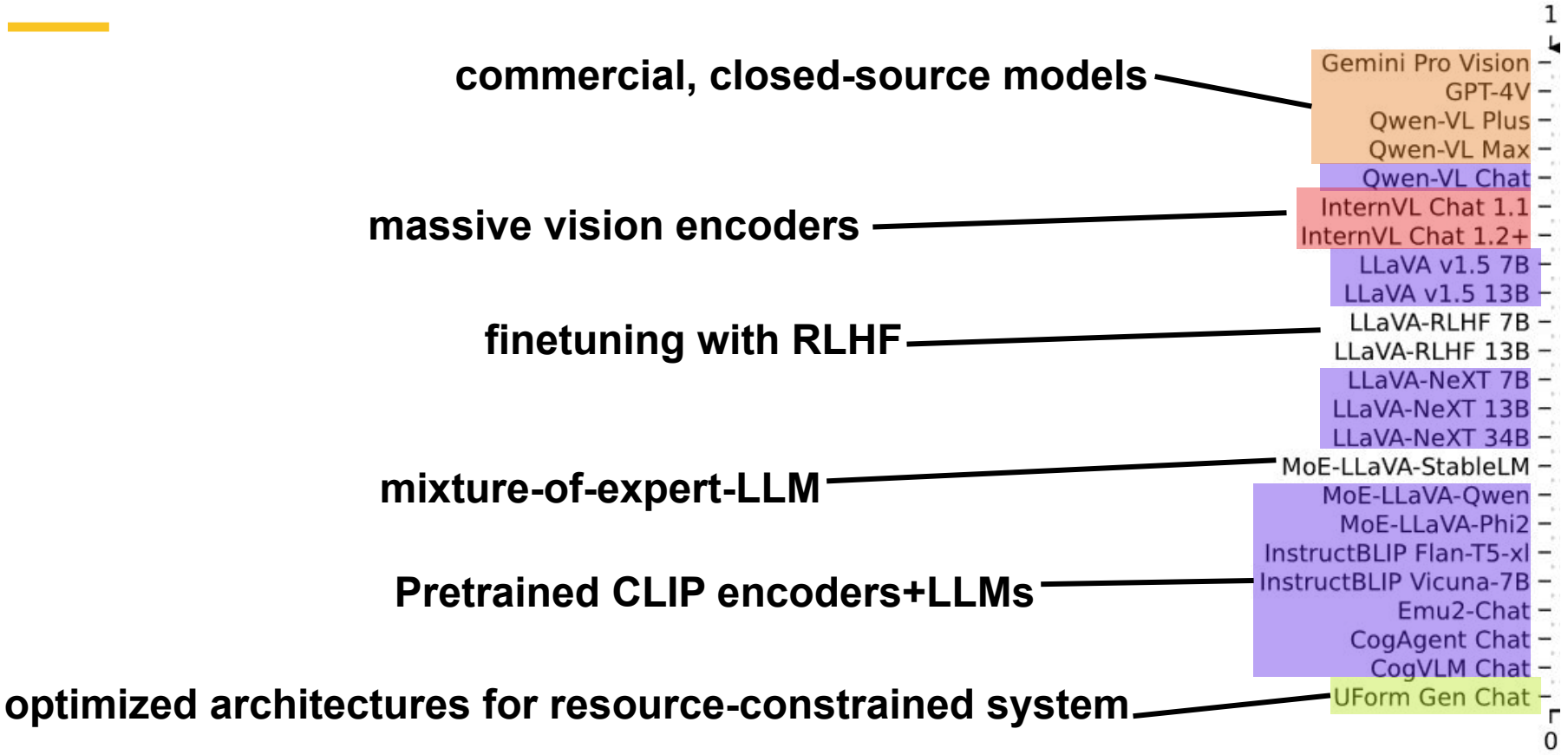
Image Captioning

Prompt
"Describe the image. Keep your response short." (default)
"Describe the image."
"Describe the image. Be precise."

Source: Table 6, Appendix C.2.

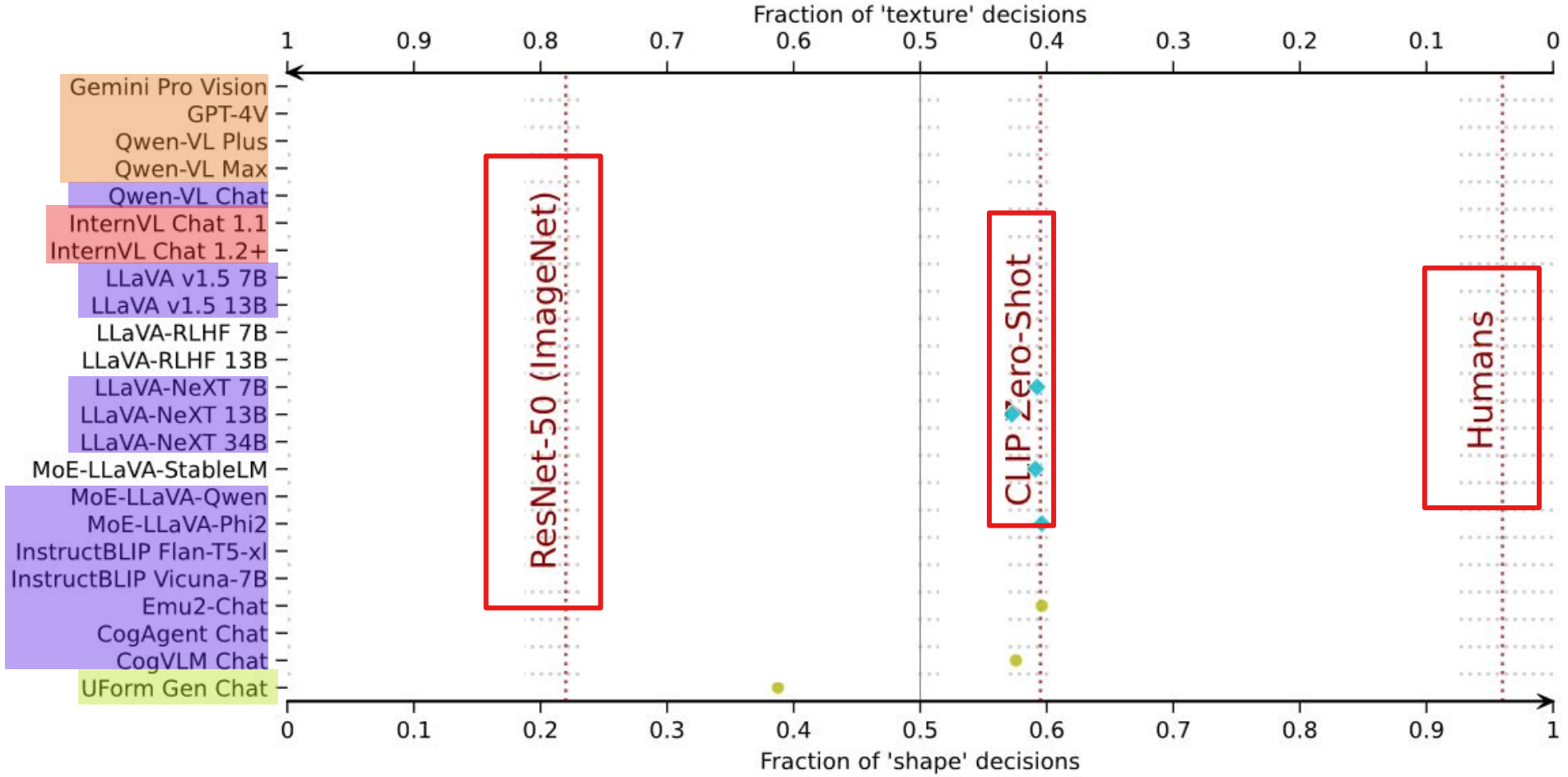
Mike Miller is next

Experiment 1 - VLM Cue Bias Analysis: *Models*



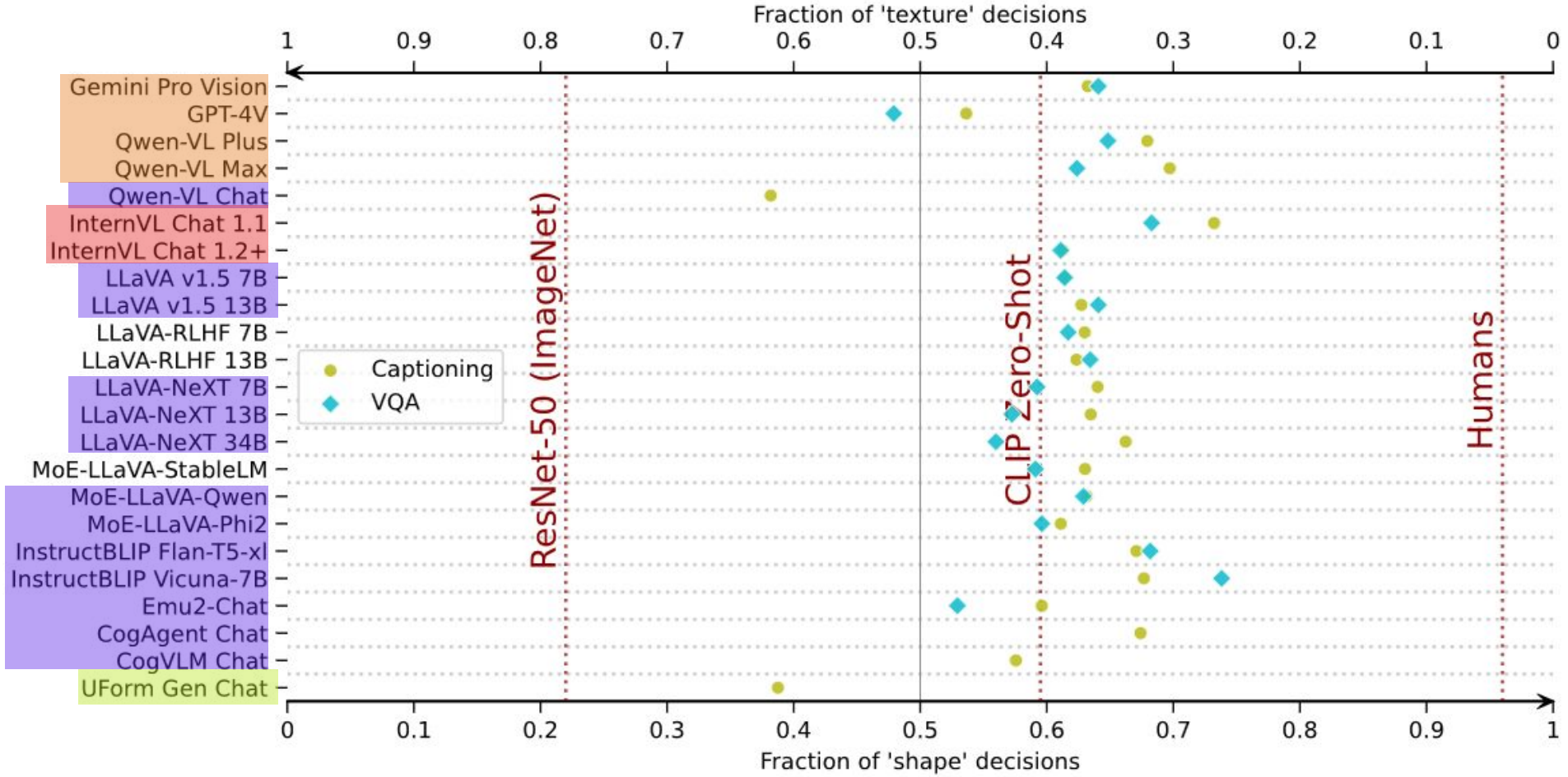
Source: Figure 2 from paper.

Experiment 1 - VLM Cue Bias Analysis: *Results*



Source: Figure 2.

Experiment 1 - VLM Cue Bias Analysis: *Results*



Source: Figure 2.

Experiment 1 - VLM Cue Bias Analysis: *Results*

Model	VQA		Image Captioning				
	Shape Bias [%]	Accuracy [%]	Shape Bias [%]	Accuracy [%]	Avg. Tokens	Single Class Ratio [%]	Generic Ratio [%]
Gemini 1.0 Pro Vision (Gemini Team, 2023)	64.1		63.2				
GPT-4V (Preview) (OpenAI, 2023)	47.9		53.6				
Qwen-VL Plus (Qwen Team, 2024)	64.8		67.9				
Qwen-VL Max (Qwen Team, 2024)	62.4		69.7				
Qwen-VL Chat (Bai et al., 2023b)	-		38.2				
InternVL Chat 1.1 (Chen et al., 2024)	68.3		73.2				
InternVL Chat 1.2+ (Chen et al., 2024)	61.1		61.3				
LLaVA v1.5 7B (Liu et al., 2023a)	61.4		61.4				
LLaVA v1.5 13B (Liu et al., 2023a)	64.1		62.7				
LLaVA-RLHF 7B (Sun et al., 2023c)	61.7		63.0				
LLaVA-RLHF 13B (Sun et al., 2023c)	63.4		62.3				
LLaVA-NeXT 7B (Liu et al., 2024)	59.2		64.0				
LLaVA-NeXT 13B (Liu et al., 2024)	57.2		63.5				
LLaVA-NeXT 34B (Liu et al., 2024)	56.0		66.2				
MoE-LLaVA-StableLM (Lin et al., 2024)	59.1		63.0				
MoE-LLaVA-Qwen (Lin et al., 2024)	62.9		63.2				
MoE-LLaVA-Phi2 (Lin et al., 2024)	59.6		61.1				
InstructBLIP Flan-T5-xl (Dai et al., 2023)	68.2		67.1				
InstructBLIP Vicuna-7B (Dai et al., 2023)	73.8		67.7				
Emu2-Chat (Sun et al., 2023a)	52.9		59.6				
CogAgent Chat (Hong et al., 2023)	-		67.4				
CogVLM Chat (Wang et al., 2023)	-		57.6				
UForm Gen Chat (Kim et al., 2023)	-		38.8				

Source: Table 4.

Experiment 1 - VLM Cue Bias Analysis: *Results*

Image Captioning Response Qualities

Model	All responses		Non-generic	
	Shape Bias [%]	Accuracy [%]	Shape Bias [%]	Accuracy [%]
Gemini 1.0 Pro Vision	63.2	68.00	65.7	88.40
GPT-4V (Preview)	53.6	52.67	62.9	93.14
Qwen-VL Plus	67.9	65.50	71.9	88.56
Qwen-VL Max	69.7	68.50	72.1	91.52
Qwen-VL Chat	38.2	67.42	40.1	83.92
InternVL Chat 1.1	73.2	75.58	74.5	87.89
InternVL Chat 1.2+	61.3	82.42	62.3	88.15
LLaVA v1.5 7B	61.4	76.08	62.8	87.24
LLaVA v1.5 13B	62.7	75.58	65.1	88.24
LLaVA-RLHF 7B	63.0	71.83	64.7	83.80
LLaVA-RLHF 13B	62.3	73.25	66.3	86.08
LLaVA-NeXT 7B	64.0	65.08	66.9	92.48
LLaVA-NeXT 13B	63.5	65.25	65.3	92.95
LLaVA-NeXT 34B	66.2	57.50	73.6	96.39
MoE-LLaVA-StableLM	63.0	73.92	64.1	86.28
MoE-LLaVA-Qwen	63.2	75.33	64.4	88.03
MoE-LLaVA-Phi2	61.1	75.42	63.1	86.05
InstructBLIP Flan-T5-xl	67.1	81.50	68.7	89.09
InstructBLIP Vicuna-7B	67.7	80.67	68.4	90.27
Emu2-Chat	59.6	65.00	60.3	89.94
CogAgent Chat	67.4	60.33	70.8	97.38
CogVLM Chat	57.6	66.58	61.9	93.72
UForm Gen Chat	38.8	64.50	37.9	83.00

All of these increased.

All but one of these did not.

Source: Table 9.

Experiment 1 - VLM Cue Bias Analysis: *Vision Encoding Biases*

Performance Comparison between VLM Models and Their Vision Encoders.

VLM	Vision Encoder	VLM - Encoder [%] Accuracy	Shape Bias	Error Consistency [%]
LLaVA v1.5 13B (Liu et al., 2023a) LLaVA v1.5 7B (Liu et al., 2023a) LLaVA-NeXT 34B (Liu et al., 2024) LLaVA-NeXT 13B (Liu et al., 2024) LLaVA-NeXT 7B (Liu et al., 2024) MoE-LLaVA v1.5 Phi2 x4 (Lin et al., 2024) MoE-LLaVA v1.5 Qwen x4 (Lin et al., 2024) MoE-LLaVA v1.5 StableLM x4 (Lin et al., 2024)	❄ CLIP ViT-L/14@336px (Radford et al., 2021)			
InstructBLIP FLAN-T5-XL (Dai et al., 2023) InstructBLIP Vicuna-7B (Dai et al., 2023)	❄ EVA-01-CLIP ViT-g/14@224px (Sun et al., 2023b)			-
Emu2-Chat (Sun et al., 2023a)	❄ EVA-02-CLIP-E/14+@448px (Sun et al., 2023b)			-

Source: Table 1.

Experiment 1 - VLM Cue Bias Analysis: *Vision Encoding Biases*

Performance Comparison between VLM Models and Their Vision Encoders.

VLM	Vision Encoder	VLM - Encoder [%] Accuracy	Shape Bias	Error Consistency [%]
LLaVA v1.5 13B (Liu et al., 2023a)	❄ CLIP ViT-L/14@336px (Radford et al., 2021)	−3.50	+4.2*	73.5
LLaVA v1.5 7B (Liu et al., 2023a)		−3.00	+1.5	76.8
LLaVA-NeXT 34B (Liu et al., 2024)		−9.92	−3.9*	67.1
LLaVA-NeXT 13B (Liu et al., 2024)		−0.33	−2.7	70.9
LLaVA-NeXT 7B (Liu et al., 2024)		−1.08	−0.2	67.5
MoE-LLaVA v1.5 Phi2 x4 (Lin et al., 2024)		−1.42	−0.3	73.4
MoE-LLaVA v1.5 Qwen x4 (Lin et al., 2024)		−24.25	+3.0	51.4
MoE-LLaVA v1.5 StableLM x4 (Lin et al., 2024)		−3.67	−0.8	73.1
InstructBLIP FLAN-T5-XL (Dai et al., 2023)	❄ EVA-01-CLIP ViT-g/14@224px (Sun et al., 2023b)	−6.83	+1.8	73.7
InstructBLIP Vicuna-7B (Dai et al., 2023)		−14.42	+7.4*	78.7
Emu2-Chat (Sun et al., 2023a)	❄ EVA-02-CLIP-E/14+@448px (Sun et al., 2023b)	−11.08	−9.5*	61.0

All VLMs were less accurate than their encoders!

Source: Table 1.

Experiment 1 - VLM Cue Bias Analysis: *Vision Encoding Biases*

Performance Comparison between VLM Models and Their Vision Encoders.

VLM	Vision Encoder	VLM - Encoder [%] Accuracy	VLM - Encoder [%] Shape Bias	Error Consistency [%]
LLaVA v1.5 13B (Liu et al., 2023a)	CLIP ViT-L/14@336px (Radford et al., 2021)	−3.50	+4.2*	73.5
LLaVA v1.5 7B (Liu et al., 2023a)		−3.00	+1.5	76.8
LLaVA-NeXT 34B (Liu et al., 2024)		−9.92	−3.9*	67.1
LLaVA-NeXT 13B (Liu et al., 2024)		−0.33	−2.7	70.9
LLaVA-NeXT 7B (Liu et al., 2024)		−1.08	−0.2	67.5
MoE-LLaVA v1.5 Phi2 x4 (Lin et al., 2024)		−1.42	−0.3	73.4
MoE-LLaVA v1.5 Qwen x4 (Lin et al., 2024)		−24.25	+3.0	51.4
MoE-LLaVA v1.5 StableLM x4 (Lin et al., 2024)		−3.67	−0.8	73.1
InstructBLIP FLAN-T5-XL (Dai et al., 2023)	EVA-01-CLIP ViT-g/14@224px (Sun et al., 2023b)	−6.83	+1.8	73.7
InstructBLIP Vicuna-7B (Dai et al., 2023)		−14.42	+7.4*	78.7
Emu2-Chat (Sun et al., 2023a)	EVA-02-CLIP-E/14+@448px (Sun et al., 2023b)	−11.08	−9.5*	61.0

*: Statistically significant changes in shape bias.

This wasn't consistently changed by the encoder!

Source: Table 1.

Experiment 1 - VLM Cue Bias Analysis: *Vision Encoding Biases*

Performance Comparison between VLM Models and Their Vision Encoders.

VLM	Vision Encoder	VLM - Encoder [%] Accuracy	Shape Bias	Error Consistency [%]
LLaVA v1.5 13B (Liu et al., 2023a)	CLIP ViT-L/14@336px (Radford et al., 2021)	−3.50	+4.2*	73.5
LLaVA v1.5 7B (Liu et al., 2023a)		−3.00	+1.5	76.8
LLaVA-NeXT 34B (Liu et al., 2024)		−9.92	−3.9*	67.1
LLaVA-NeXT 13B (Liu et al., 2024)		−0.33	−2.7	70.9
LLaVA-NeXT 7B (Liu et al., 2024)		−1.08	−0.2	67.5
MoE-LLaVA v1.5 Phi2 x4 (Lin et al., 2024)		−1.42	−0.3	73.4
MoE-LLaVA v1.5 Qwen x4 (Lin et al., 2024)		−24.25	+3.0	51.4
MoE-LLaVA v1.5 StableLM x4 (Lin et al., 2024)		−3.67	−0.8	73.1
InstructBLIP FLAN-T5-XL (Dai et al., 2023)	EVA-01-CLIP ViT-g/14@224px (Sun et al., 2023b)	−6.83	+1.8	73.7
InstructBLIP Vicuna-7B (Dai et al., 2023)		−14.42	+7.4*	78.7
Emu2-Chat (Sun et al., 2023a)	EVA-02-CLIP-E/14+@448px (Sun et al., 2023b)	−11.08	−9.5*	61.0

*: Statistically significant changes in shape bias.

The two systems do NOT make the same errors for the same reasons!

Source: Table 1.

Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*

```
"{VQA_INSTRUCTION}
A. airplane
B. bear
C. bicycle
…
P. truck
Answer with the option’s letter from the given
choices directly."
```

Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*

🤖: {
bear: 93.82%
plane: 6.50%
...
bottles: 0.01%
...
}



Only 17.7% of the top-2 pairs contained both shape and texture.

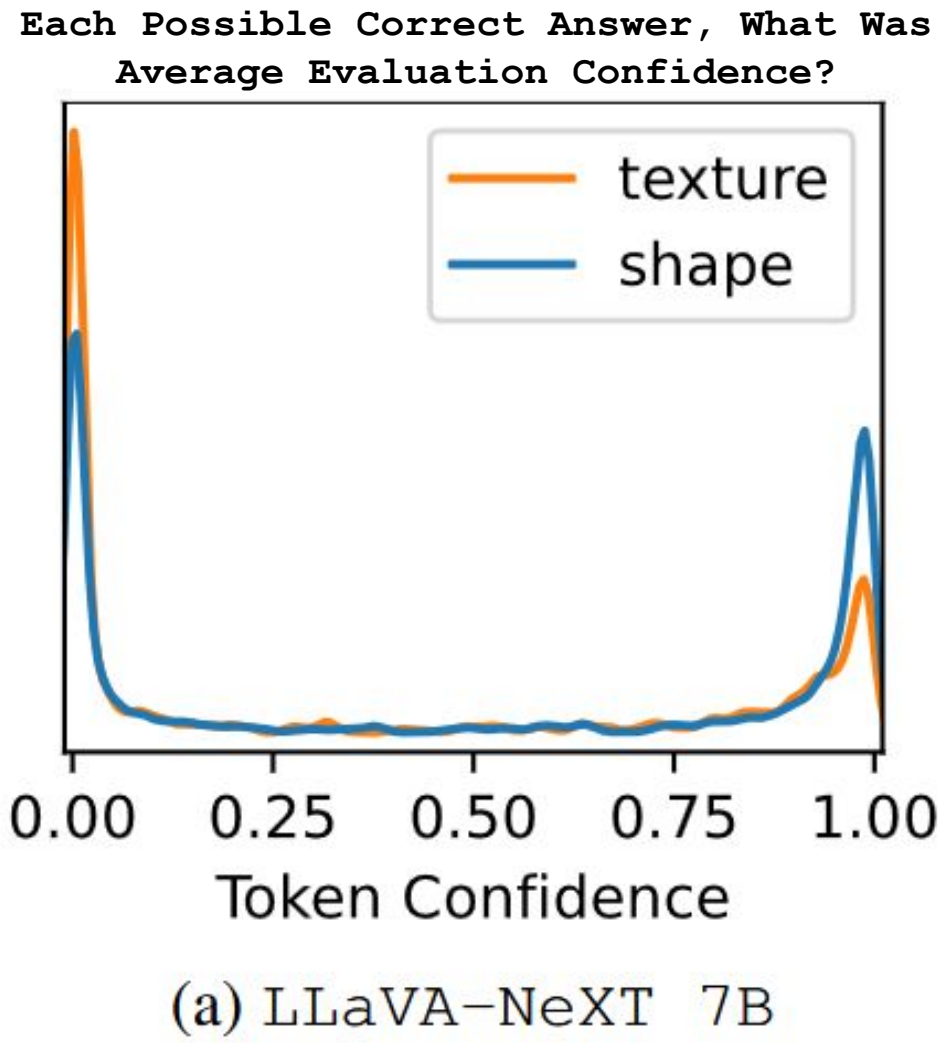
Source: Figure 3.

Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*



🤖: {
bear: 93.82%
plane: 6.50%
...
bottles: 0.01%
...
}

Source: Figure 3.

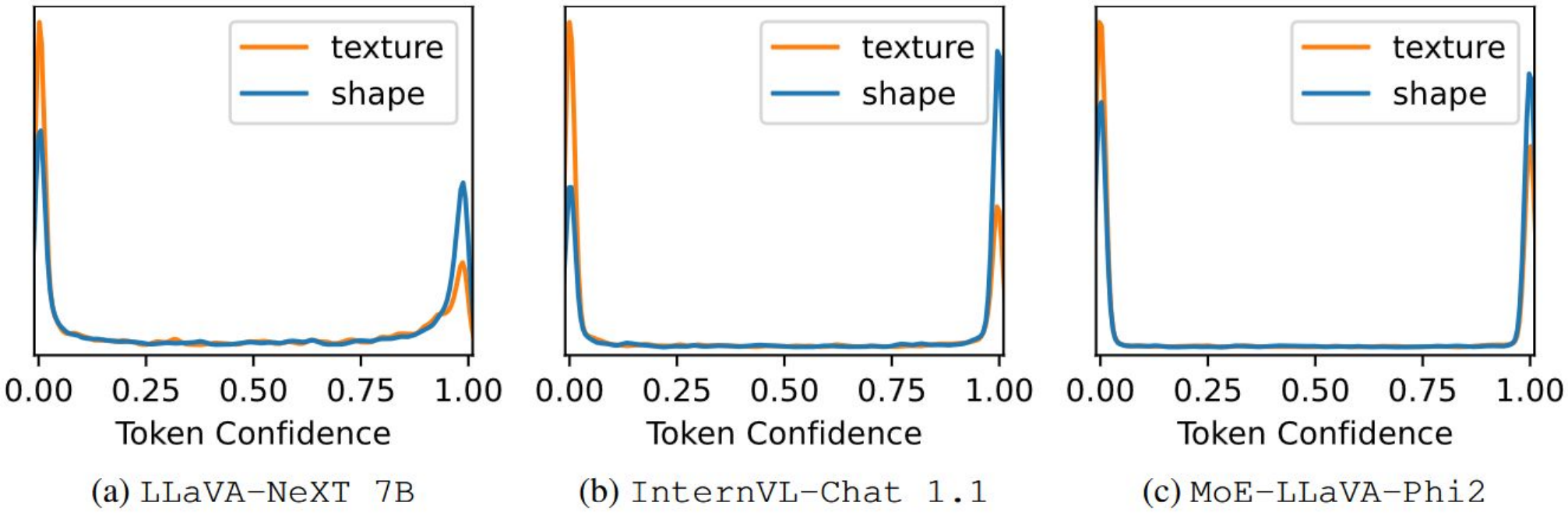


Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*



🤖: {
bear: 93.82%
plane: 6.50%
...
bottles: 0.01%
...
}

For Each Possible Correct Answer, What Was the Average Evaluation Confidence?



Source: Figure 3.

Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*

🤖1: `plane!`
🤖2: `plane!`
🤖3: `plane!`
💁‍♀️1: Cat!
💁‍♀️2: Truck!
💁‍♀️3: Cat!



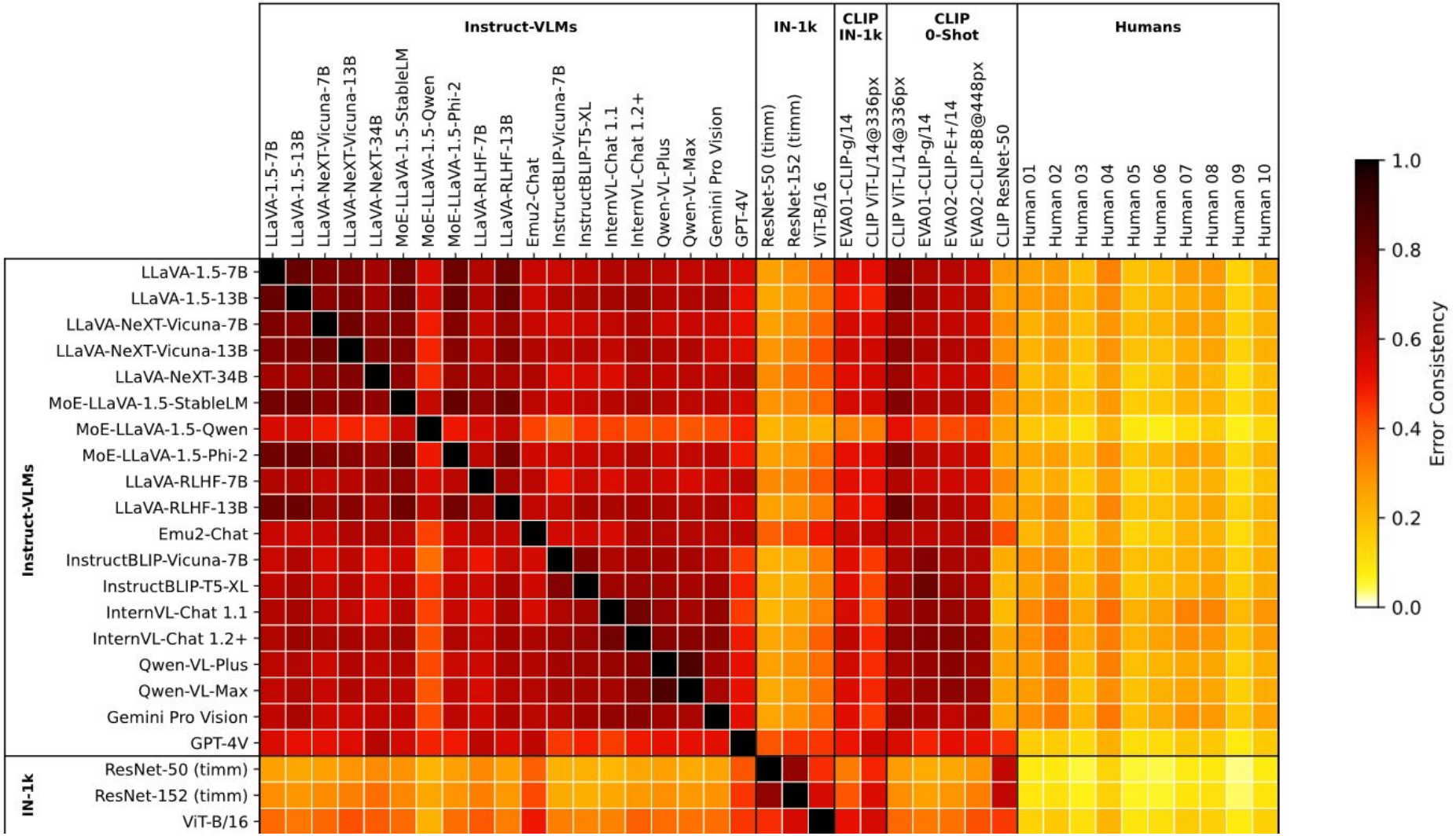
Source: Figure 13.

Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*



🤖1: `plane!`
🤖2: `plane!`
🤖3: `plane!`
👩1: Cat!
👩2: Truck!
👩3: Cat!

Source: Figure 13.

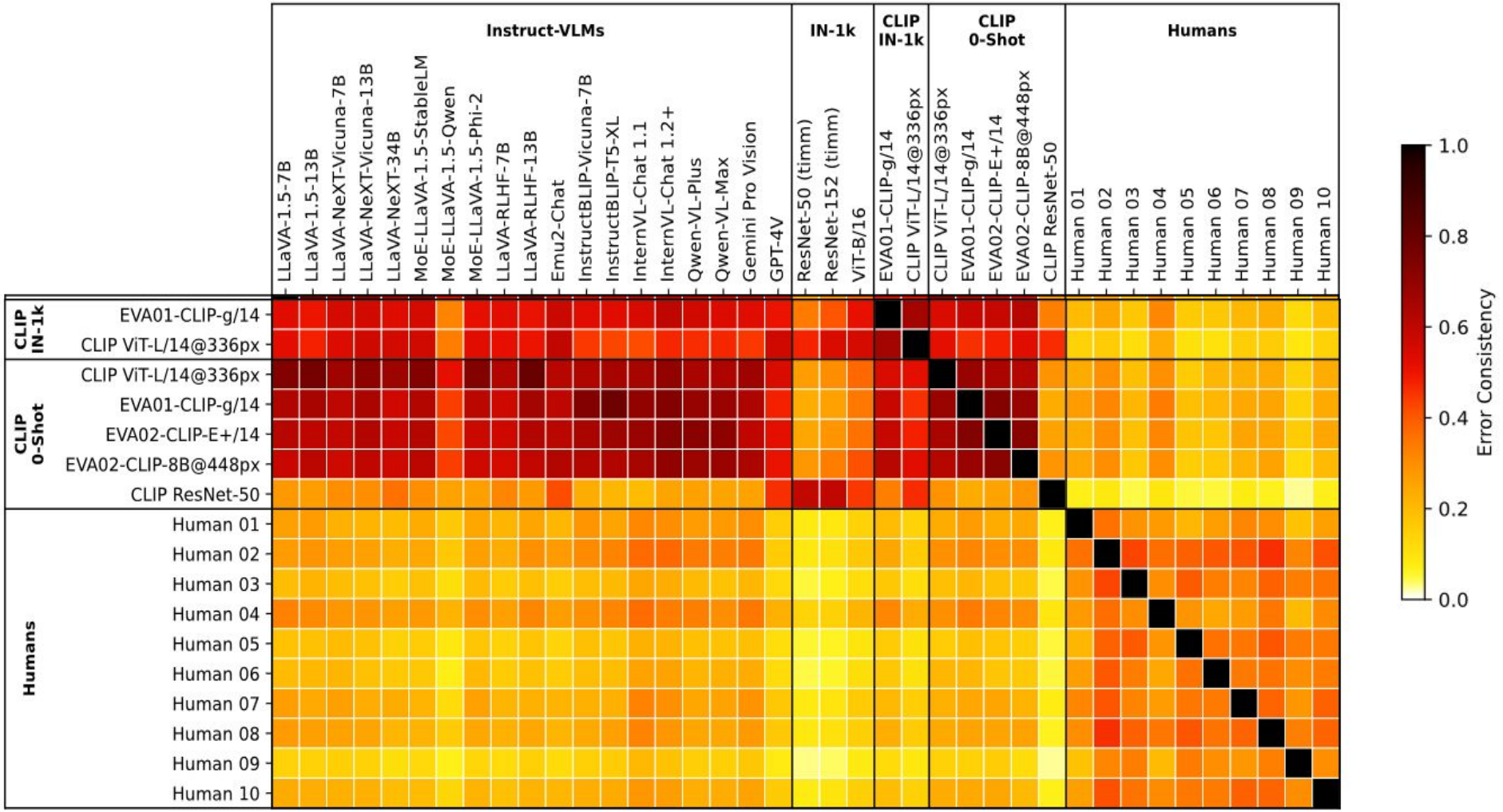


Experiment 1 - VLM Cue Bias Analysis: *LLM Processing Biases*



🤖1: plane!
🤖2: plane!
🤖3: plane!
👩1: Cat!
👩2: Truck!
👩3: Cat!

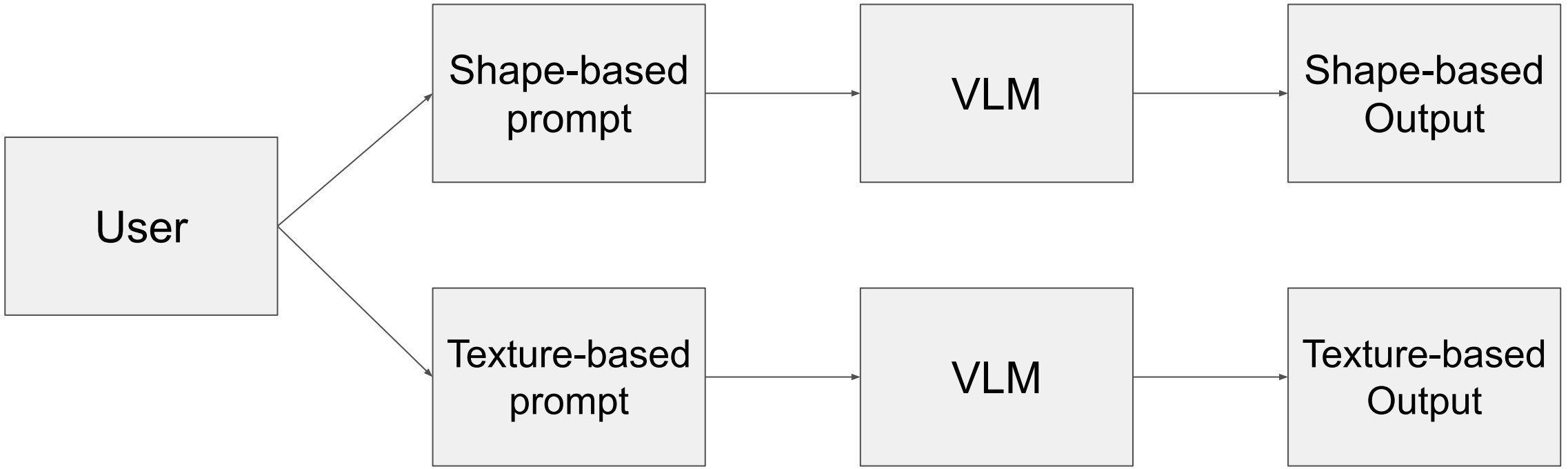
Source: Figure 13.



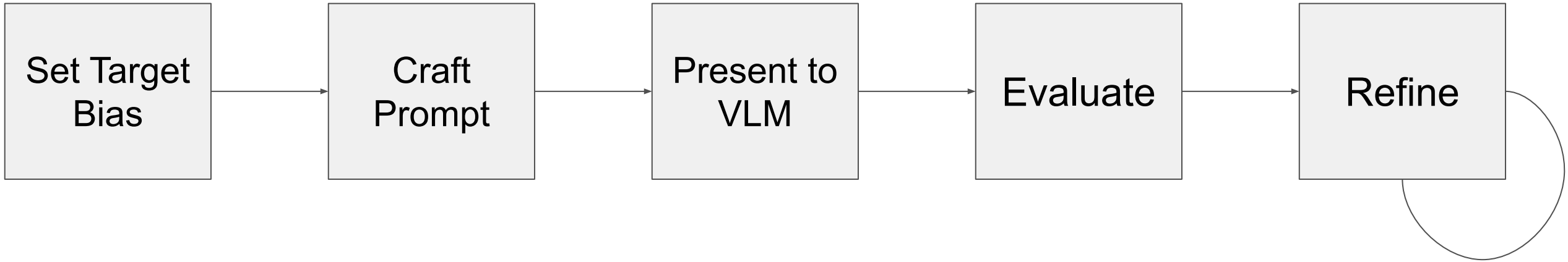
Knowing that the LLM is what causes our bias to skew towards texture…

Can We Steer Models' Bias Towards Humans' Bias Through Language?

Methodology - Active Prompt Steering (Overview)



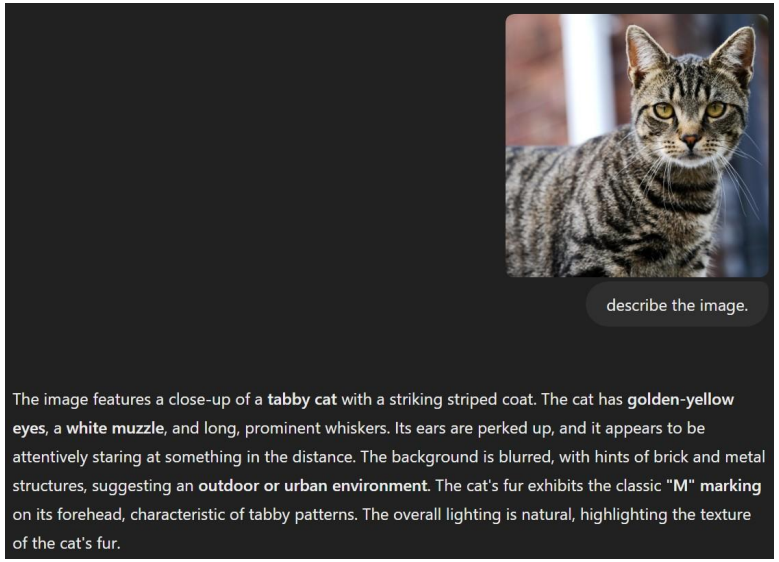
Methodology - Active Prompt Steering (Steps)

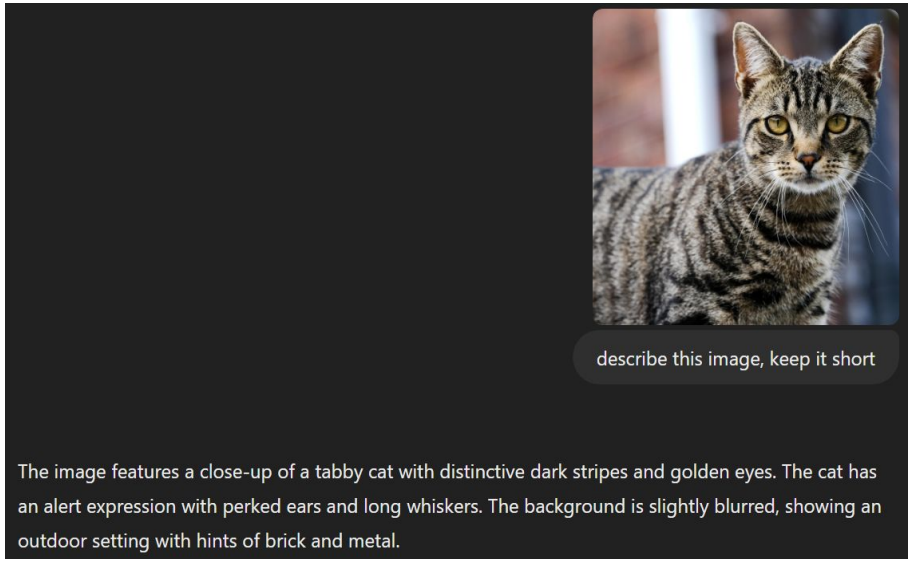


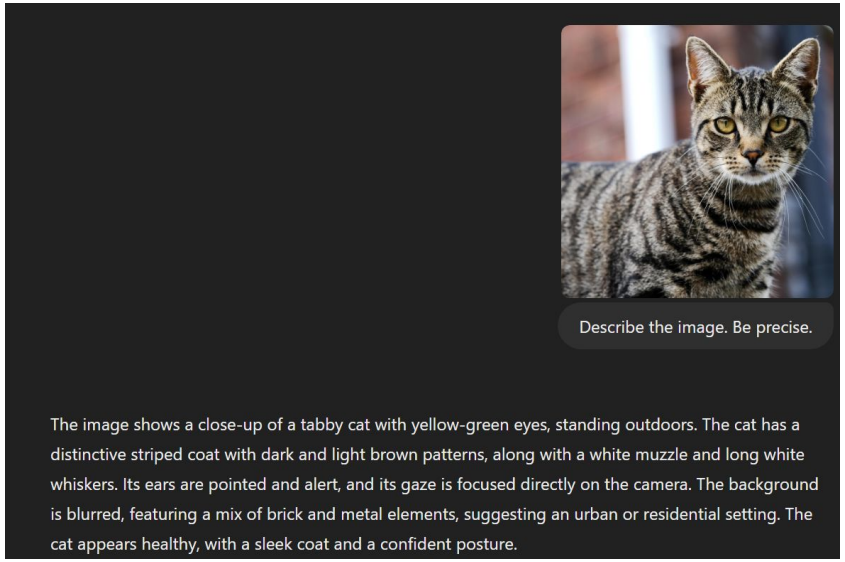
Methodology - Image Captioning (Overview)

- Open-Ended Task: “Describe the image in a short sentence.”
- Challenge: Can mention shape, texture, both, or neither
- Mapping to Classes: Compare caption text to shape/texture labels

Source: Table 6.





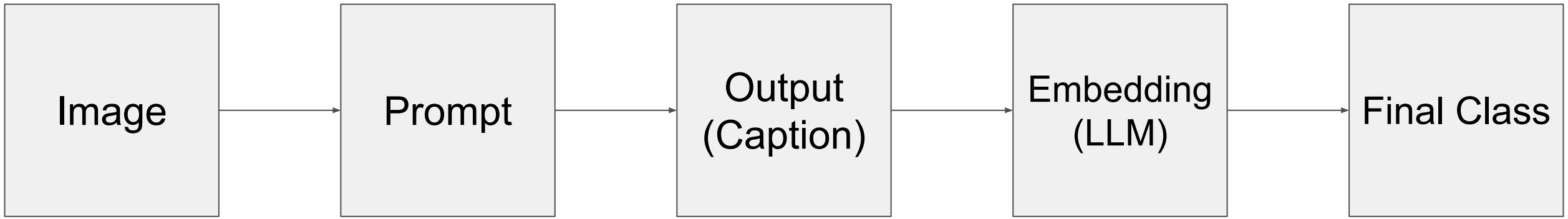


Images prompted with o1

Table 6: Exploration of alternative Image Captioning prompts.

Prompt	Shape Bias [%]	Accuracy [%]	Avg. Tokens	Generic Ratio [%]
`"Describe the image. Keep your response short."` (default)	64.0	65.08	55.5	39.5
`"Describe the image."`	63.6	68.25	202.9	46.8
`"Describe the image. Be precise."`	67.3	64.50	166.2	50.6

Methodology - Image Captioning (Detailed Steps)







```
Caption: A striped cat-like figure with gray, rough skin.

Similarity scores:
cat        score: 0.7205
elephant   score: 0.6503
zebra      score: 0.5152
dog        score: 0.4921
truck      score: 0.1508
bottle     score: 0.1287
```

```
Best shape match: cat (score=0.7301)
Best texture match: rough skin (score=0.6605)

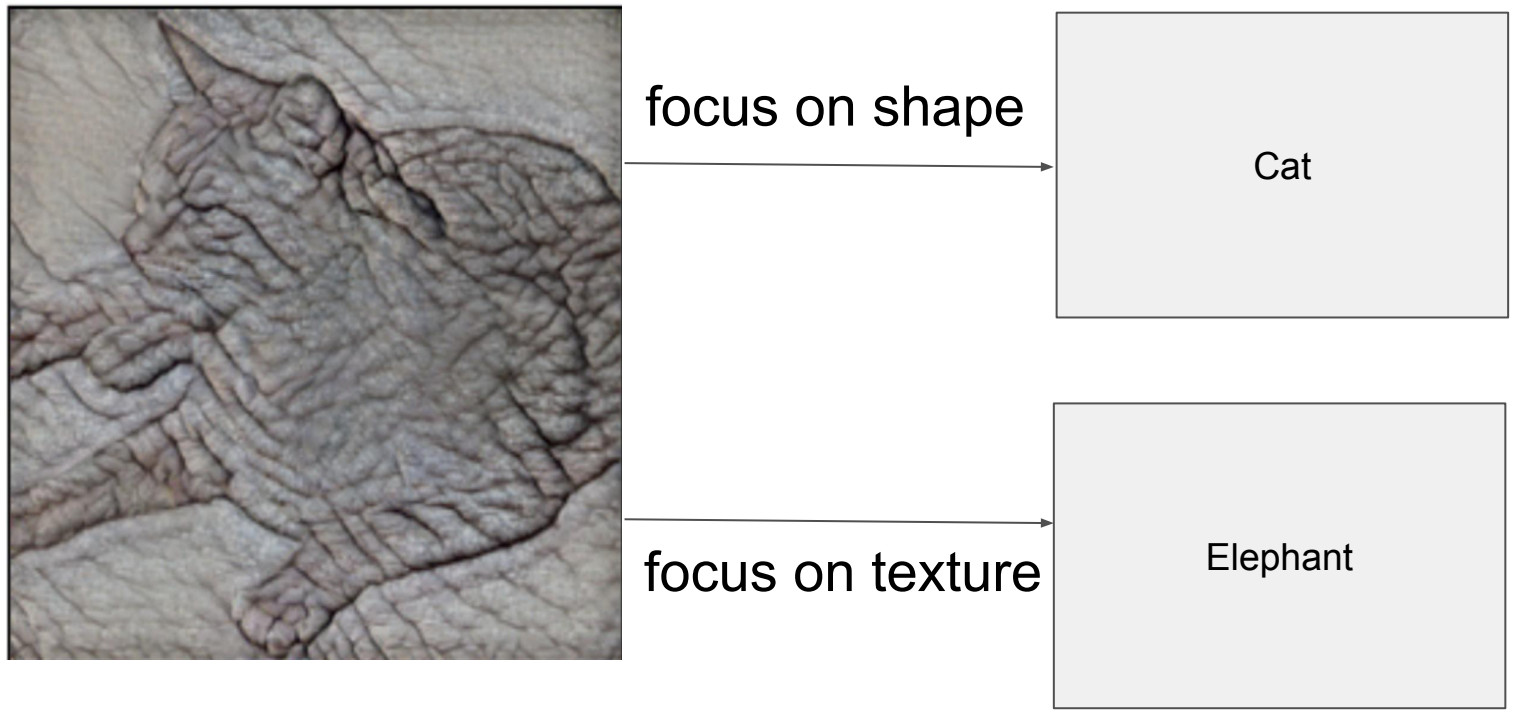
Implied cue: SHAPE → 'cat' with score=0.7301
```

Experiment 2: Prompt-Based Steering (Setup)

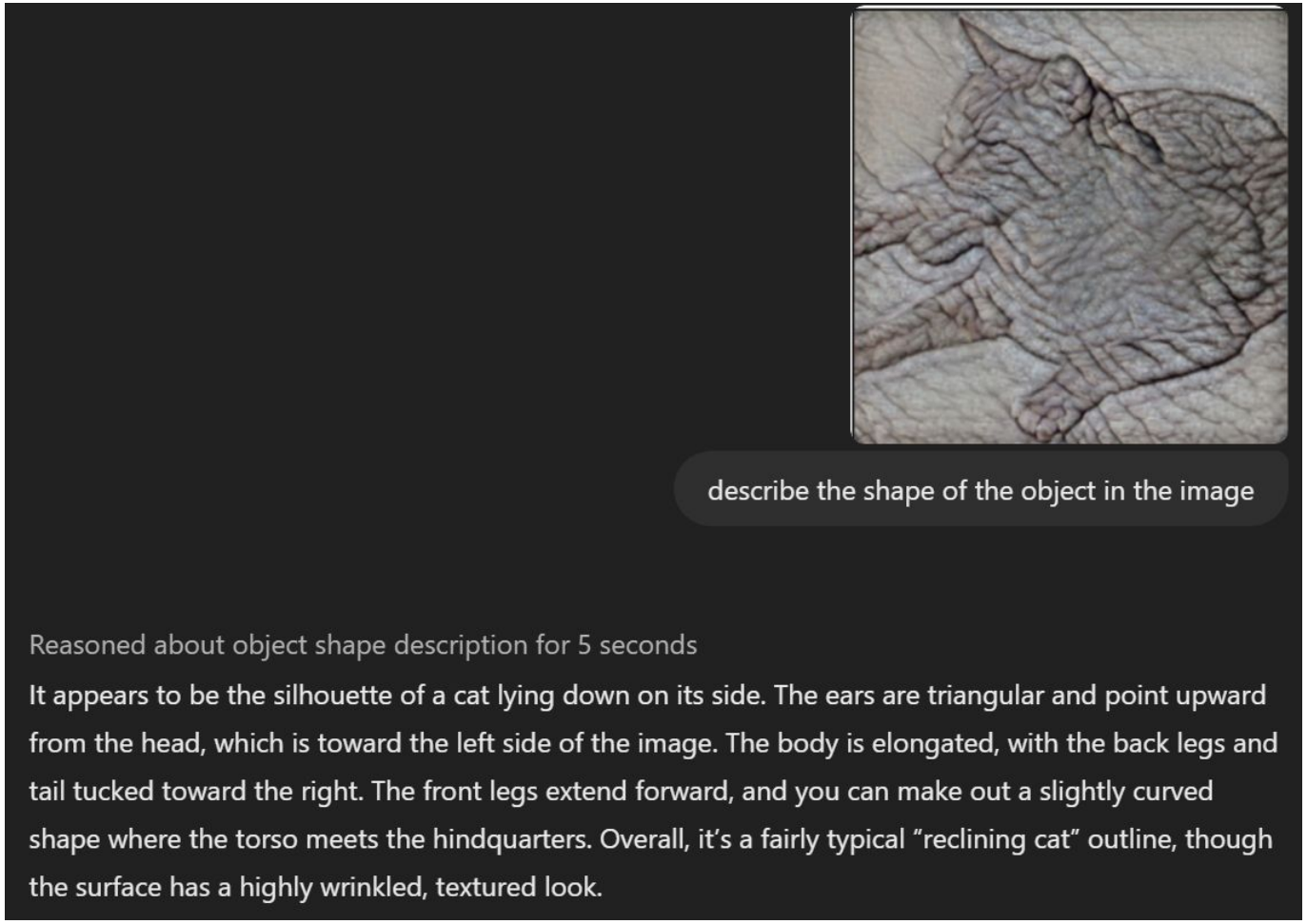
- Context: Building from Experiment 1's bias measurement
- Objective: Actively shift shape vs. texture reliance

Neutral Prompt	Shape-Based	Texture-Based
"Which option best describes the image?"	"Which option best describes the shape?"	"Focus on the texture or surface pattern…"

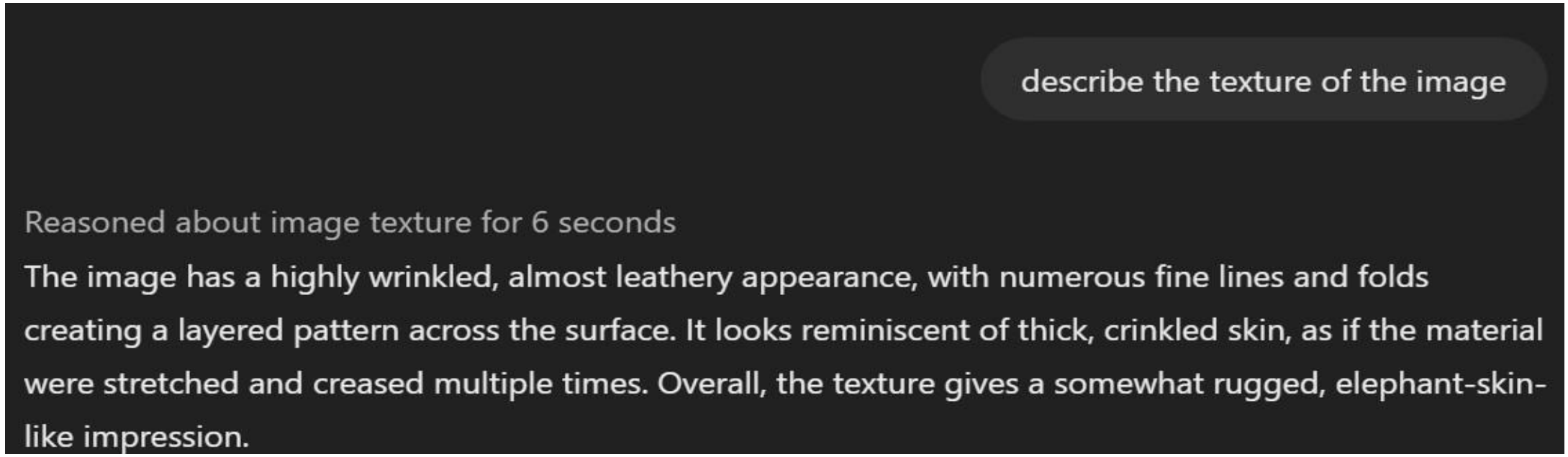
Experiment 2: Steering Toward Shape & Texture (Process)



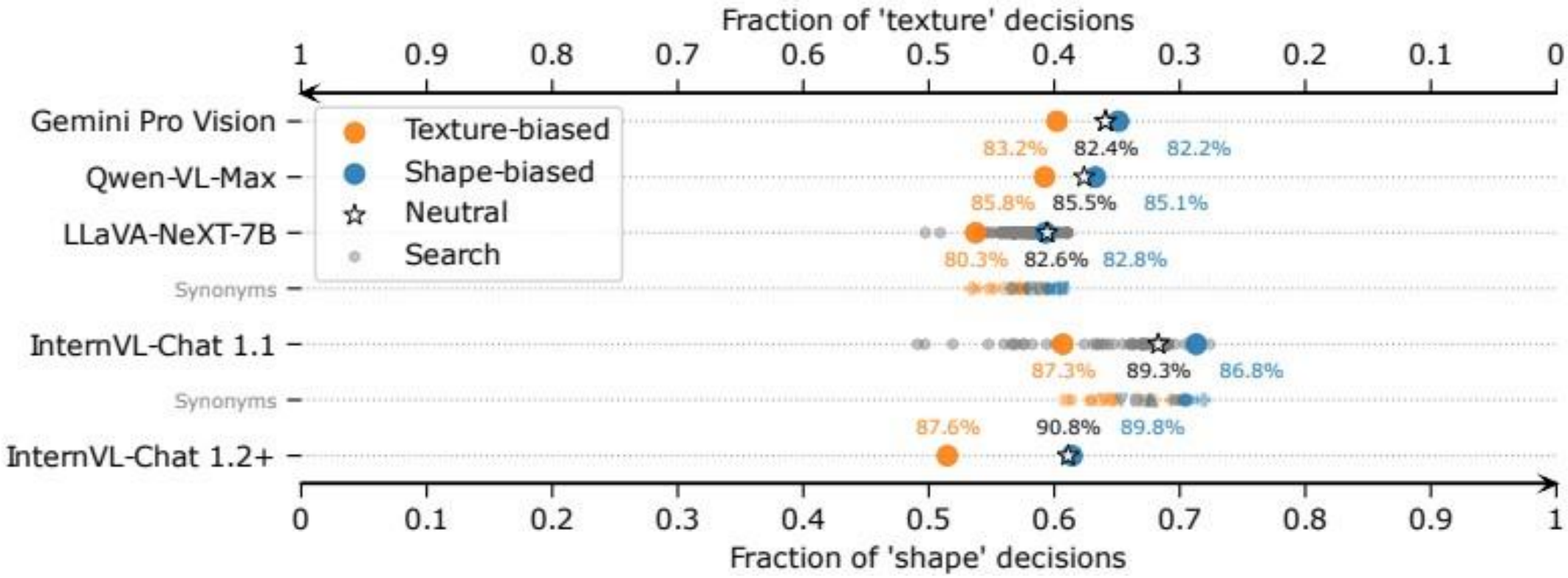
Experiment 2: Example Outputs (Shape vs. Texture)



Experiment 2: Example Outputs (Shape vs. Texture)



Experiment 2: Steering Toward Shape & Texture (Results)



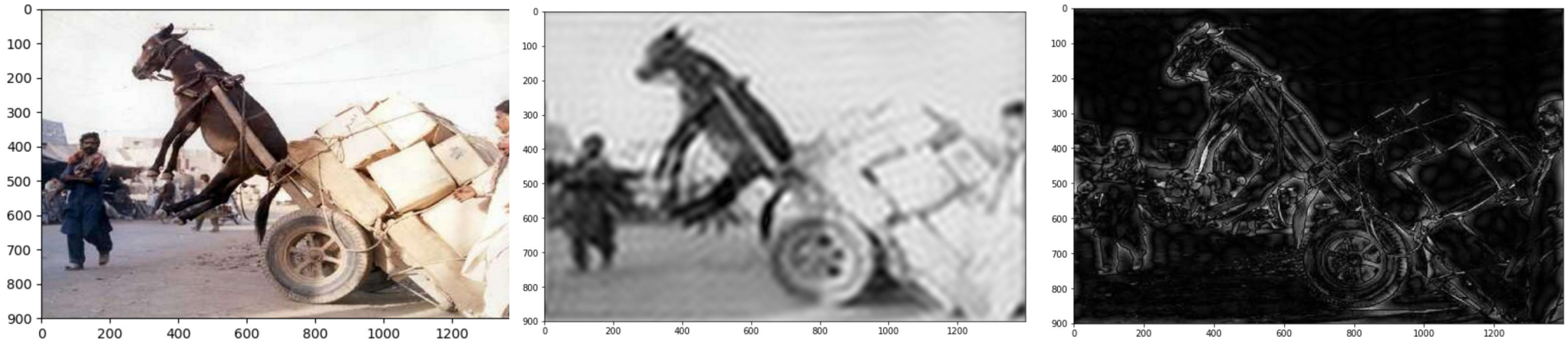
Source: Figure 4 from paper.

Knowing that language can be used to steer towards shape cues…

What about other visual cues?

Low-Frequency- and High-Frequency-Biased Outputs

- What is it?
- slow and smooth vs fast and sharp changes in intensity.



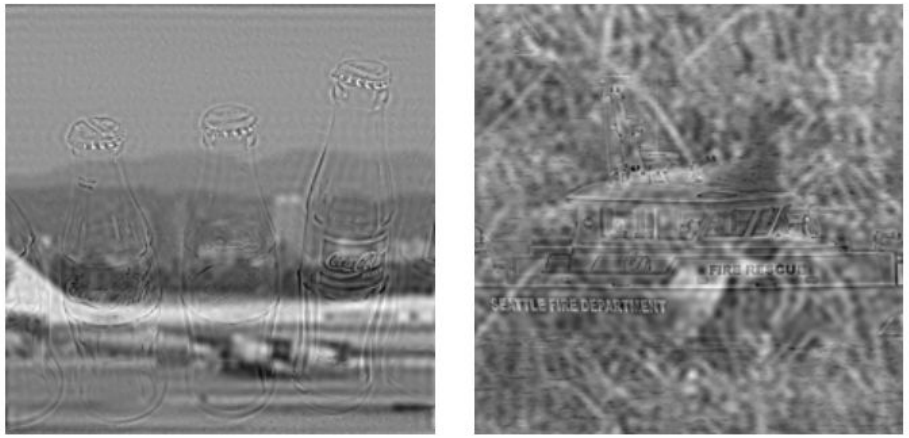
from Week 2: Images in Frequency Domain

Experiment 3 - Steering Towards *Low-, High-Frequency-Biased* Outputs

- Study the capacity of steering for other biases.
- What about low and high frequency components.
- Dataset:
  - 1200 samples
  - 16 ImageNet super classes
  - Blend of HF and LF components

Experiment 3 - Steering Towards *Low-, High-Frequency-Biased* Outputs

- Neutral prompt and Automated prompt search
- Default bias is LF Bias
- 2 Models were tested:
  - LLaVA-NeXT 7B
  - InternVL-Chat 1.1 19B



Source: Figure 9(b) from the paper.

LF:	airplane	bird
HF:	bottle	boat

Experiment 3 - Steering Towards *Low-, High-Frequency-Biased* Outputs

- Did not significantly change Accuracy
- Two statistically significant changes in bias from steering.
- Note balance in bias among models.

Model	Prompt	Accuracy [%]	LF Bias [%]
InternVL-Chat 1.1	neutral	92.92	34.5
	search (max)	90.33	38.6*
	search (min)	91.33	32.9
LLaVA-NeXT 7B	neutral	82.83	52.4
	search (max)	84.25	54.5
	search (min)	82.67	48.7*

Source: Table 2.

Comparing Results to Human Biases

- Models still lag behind human shape bias.
- 96% vs little over 70%
- Humans tend to have a low-frequency bias

We will all work on these asynchronously and meet on Monday morning to finalize them

The Limitations of Steerability and the Study

- Steerability
  - Vision steering affects model bias higher than Language steering
  - Steering usually negatively affects accuracy
  - Increasing Texture vs Shape bias
- Study
  - Unexpected behavior (like refusal to answer)
  - Unlimited alternative prompts

Discussion and Other Limitations of the Study

-

Strengths

- Paper opens the ability to change what aspects of an image a model focuses on without having to retrain or rewrite the code unlike say a deep Neural Network.
- The ability to prompt steer can be applied to different LLMs, and non-technical people can take advantage of this.
- Shows the flexible and versatile nature of LLMs to possibly perform many sort of computer vision tasks or just deep learning tasks possibly involving audio or other media formats in the future.

Weaknesses

- Using language prompts is not as precise as hard-coding different forms of bias
- Might not be as expandable to radically different LLM architectures
- Possibly beyond the scope of the paper, but could explore the influence of prompt steering on other aspects beyond shape/texture bias.
- Could discuss other ways to do text/prompt steering in order to discuss the greater implications of this paper. Right now for some readers, the wider implications could be lost.

Conclusion

- Humans are biased shape over texture
- VLMs require more alignment towards human biases
- Vision encoders have a bias, but give *cue-flexible tokens* to the LLM
- LLMs drive most of the bias, can be steered through prompting
- We Can Talk Models Into Seeing The World Differently!

Future Work Ideas

- This paper focused on VLM bias *steerability* through LLM prompts, but more biases could be evaluated:
  - Color
  - Light Intensity
  - Spectral (amplitude vs. phase)
  - Spatial
  - Different Modalities

REFERENCES

- Gavrikov, Paul, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. 2025. “Can We Talk Models Into Seeing the World Differently?” In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=iVMcYxTiVM.
- Robert Geirhos, et al. 2019. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. https://arxiv.org/abs/1811.12231
- https://github.com/rgeirhos/texture-vs-shape/blob/master/stimuli/style-transfer-preprocessed-512/airplane/airplane4-dog2.png